The University Of Sheffield.

Gene Expression Analysis.
Use of Bioconductor and *limma*

The University Of Sheffield.

Today's Outline

Part A : Introduction of probabilistic models
Bayesian Inference
Probabilistic models for gene expression data
microarray data: Gene Expression Estimation with *puma*

Break – 10 min (questions if any)

Part B : Probabilistic models for gene expression data
RNA-Seq data: bitSeq
Differential Expression Analysis
False Discovery rate

Bioconductor and *limma*

The University Of Sheffield.

Today we will be learning:

- What are probabilistic models and how they work on high throughput data
- Bayes' Rule
- Uncertainty in probabilistic models
- *puma:* probabilistic model for gene expression data
- Use of probabilistic models for RNA-Seq
- Differential Expression Analysis for gene expression data
- The concept of False Discovery Rate applied to gene expression data
- Bioconductor : Open Source Platform for gene expression data
- *limma* package

The University Of Sheffield.

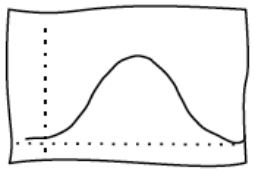
Probabilistic Models

A *probability model* is a mathematical representation of a random phenomenon. It is defined by its sample space, events within the sample space, and probabilities associated with each event.

These are models that represents unknowns in terms of probability distributions instead of values and a confidence interval.

For example if we assume that the data is generated by:

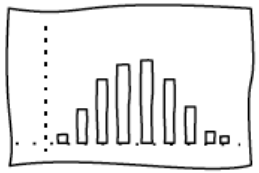
$$\theta \sim N(\mu|\sigma)$$



I

we can measure few samples of it, but we don't know the true distribution:

$$\theta \sim \mathbf{S} = \{s_1, \ldots, s_n\}$$



The University Of Sheffield.

Probabilistic Models (cont.)

We can use a probability theory to manipulate those functions (probabilities) and make inference on the unknown parameters as well as evaluate the *uncertainty* that is associated to their estimates. In other words:

- We describes data that one could observe from a system
- We use the mathematics of probability theory to express all forms of uncertainty and noise associated with our model...
- We use inverse probability (i.e. Bayes rule) to infer unknown quantities, adapt our models, make predictions and **learn** from data.

Why are they good?

- They Faithfully represent uncertainty in our model structure and parameters and noise in our data
- They are automated, adaptive and robust
- They scale well to large data sets

The University Of Sheffield.

Bayes' Rule

$$P(hypothesis \mid data) = \frac{P(data \mid hypothesis)P(hypothesis)}{P(data)}$$

Rev'd Thomas Bayes (1702-1761)

- Bayes rule tells us how to do predict outcomes of hypotheses from data. We can do inference about hypotheses once we have observed the data

- Learning and prediction can be seen as forms of inference. If we use Bayes' Rule we call it *Bayesian Inference*

P(hypothesis) = prior
P(hypothesis|data) = posterior
P(data|hypothesis) = likelihood
P(data) = marginal likelihood

Not always possible to be computational efficient and likelihood are estimated using sampling methods.

The University Of Sheffield.

Uncertainty in Biology

In biology the complexity of the systems is such that we cannot measure everything and predictions of data required an additional “knowledge” to become meaningful.

This “knowledge” needs to be quantified in a way that reflects our prior knowledge of the systems and what we were able to measure (observe) and adjusted once the measurements are computed. It opens the way to quantifying uncertainty.

The evolution of the technology for biological sciences enables us to apply the concepts of uncertainty on complex biological data.

When we cannot quantify the parameters that are needed to describe the system we have to include the uncertainty associated to this “unknown” in our model.

We use probabilistic models to build machines (models implemented in an algorithm) that *learn* from the data. This filed of research is called **Machine Learning**.

The University Of Sheffield.

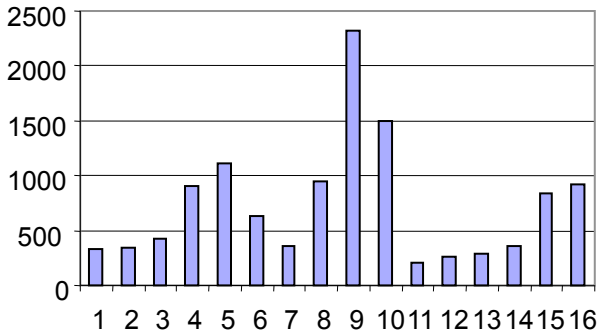
How do probabilistic model help in gene expression analysis?

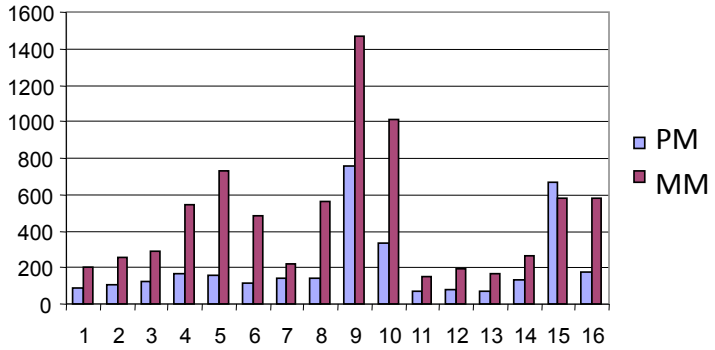
We know that:

1. Summarise to a single expression level the probe intensities for each probe set

2. Estimate the variations introduced by
   background effect
   probe affinity effect

3. Some PM/MM pairs are more reliable than others





4. The signal needs to be scaled before comparing data from different arrays

They can help to define a measure that best represent the absolute expression level of each gene on the chip?

The University Of Sheffield.

The approaches

Use **single point statistics**

make use of the information we have to define values that estimate gene expression

MAS 5.0
RMA – GCRMA

Use a **probabilistic approach**

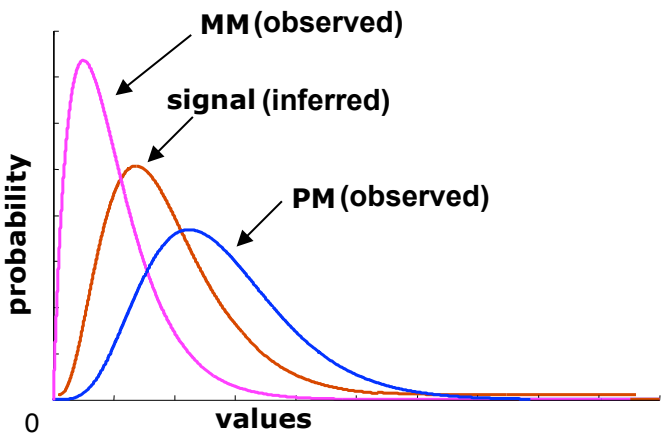
make use of the observed data to estimate probability functions that have generated that data

built a noise model and evaluate the uncertainty in the system

Estimates of gene expression will be the most probable value that summarises the observed values.

The University Of Sheffield.

puma
Propagating Uncertainty in Microarray Analysis



Milo M *et al*, Biochem transcation 2003
Liu X *et al*, Bioinformatics 2005
***Pearson R et al*, BMC Bioinformatics, 2009**

Estimate the distribution of the data and we learn the parameters to define it from the data (gamma distributions)

We built priors on the hypothesis (our belief is that the true signal is gamma distributed

We then calculate the likelihood using the model defined by Affymetrix

Signal= PM-MM

We apply Bayes' rule to calculate the signal distribution (posterior)

Computational Efficient --- we don't need to use sampling methods.

The University Of Sheffield.

The advantage of using uncertainty

Robust gene expression estimates depend on how well we are able to quantify the uncertainty.

Down stream analysis will be more effective and number of false positives will be reduced.

Probability of Positive Log Ratio | one sided Bayesian test in which we use also the variances

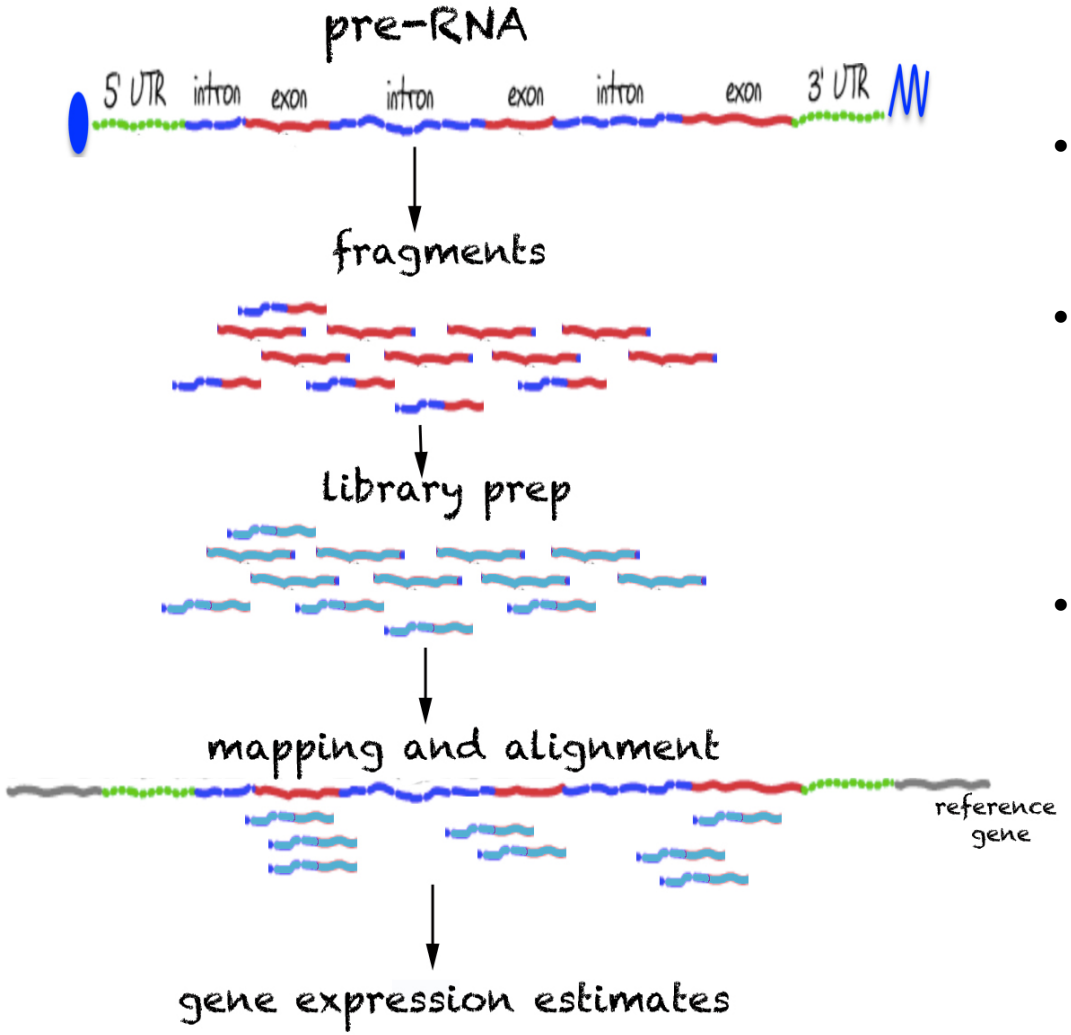
The University Of Sheffield.

The advantage of using uncertainty (cont.)

Improves accuracy in High level Analysis:

- DE Analysis using a ranking based on Probabilities of Positive Log Ratio
  - Ranking – defining False Discovery Rate (FDR) and q-values using PPLR
- Principal component analysis (*pumaPCA*)
- Clustering methods with mixture components (*pumaClust*)
- Linking SNPs and gene expression to identify functional effects. (*NG*)
- Gene networks – we are working on this.

Probabilistic methods for RNA-Sequencing



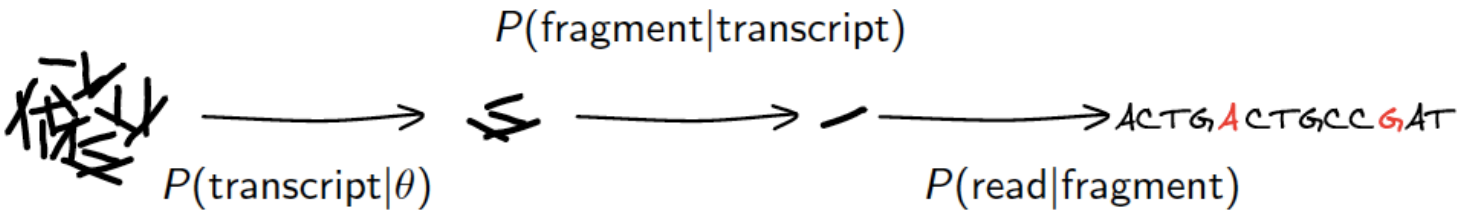
- We can use the Bayes' Rule to build a model based on this workflow
- Calculate the gene expression estimates by identifying the most probable reads associated to the mapped genes and count them
- We can do this for all the isoforms allowing for multiple matches when reads are aligned to transcriptome. Increase complexity but important

Glaus P., Honkela A., and Rattray M. (2012). Bioinformatics, 28(13),1721-1728.
Package: Bioconductor 2.10 bitSEq

The University Of Sheffield.

RNA-Seq: bitSeq approach (cont.)

Estimate the unknown relative expression of transcripts' fragments



$$P(\text{read}|\theta) = P(\text{transcript}|\theta)P(\text{fragment}|\text{transcript})P(\text{read}|\text{fragment})$$

$$P(\text{Data}|\theta) = \prod_{i=1}^{n} P(\text{read}_i|\theta)$$

$$P(\theta|\text{Data}) = \frac{P(\text{Data}|\theta)P(\theta)}{P(\text{Data})}$$

Not very efficient computationally:
To do Bayesian inference bitSeq uses Markov Chain Monte Carlo (MCMC) algorithm to produce samples from posterior. Now a new version with Variational Bayes Inference

The University Of Sheffield.

RNA-Seq: bitSeq approach (cont.)

Propagating uncertainty in DE

- For transcript *m* we want to know the probability of the expression in two experiments being different
- Remove noise
- Probability of Positive Log Ratio | one sided Bayesian test in which we use also the variances
- PPLR close to 1/0 indicates confident up/down regulation

The University Of Sheffield.

Part B

The University Of Sheffield.

High Level Analysis: workflow example

1. *Visualisation of the data*

2. *High level summary*: Combine expression from replicated arrays
   - Combine expression and uncertainty (puma)
   - Combine information from replicates with single poing statistics

3. *Differential Expression Analysis:* Determine differential expression between conditions, or between more complex contrasts such as interaction terms

4. *Dimensionality reduction* – Principal Component Analysis (Week 5)

5. *Data Clustering:* Cluster data taking the expression-level uncertainty into account (week 5)

The University Of Sheffield.

Differential expression Analysis

GOAL:
Identify the most differentially expressed genes across different conditions or cases and controls.

HOW:

identify a threshold that "define" differential expression

FC values

p-values

What happens if the sample size is small?

- The fold-change becomes very sensitive to outliers
- The t-test becomes very sensitive to small variances

The University Of Sheffield.

The Fold Change

Given two gene expression values **x** and **y** the fold change is defined as

$$FC = \frac{x}{y}$$

Given two vectors $\mathbf{x}_j$ and $\mathbf{y}_j$ of gene expression measurements for controls and cases for GENE j, the fold change is defined as

$$FC_j = \frac{x_j}{y_j}$$

It can also appear as a difference when we use the log transformation of the data.

Problem: How do we manage replicates? We need to combine the data

The University Of Sheffield.

High Level summary

High level summary: Combine expression from replicated arrays

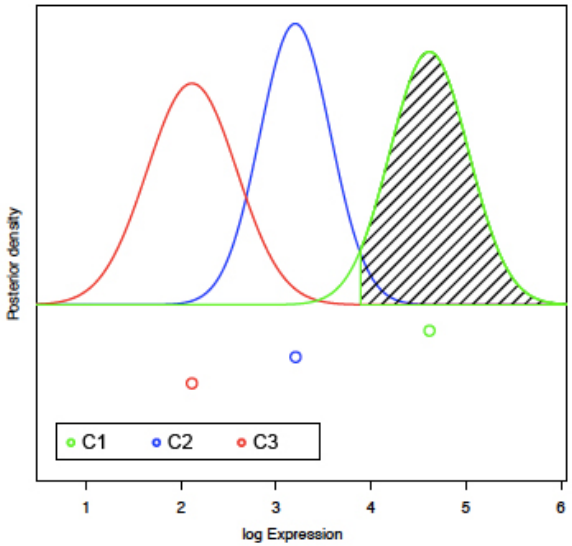
This is prior our DE and it is useful to have a representative gene expression value for the class you are studying : Wild Type vs Mutant; Disease vs Control.

- Combine information from replicates with single point statistics
- Combine expression and uncertainty (puma) using Bayesian Inference

In both cases you need a measure of uncertainty.

How do we get it in both cases?

Differential Expression: pumaDE and PPLR

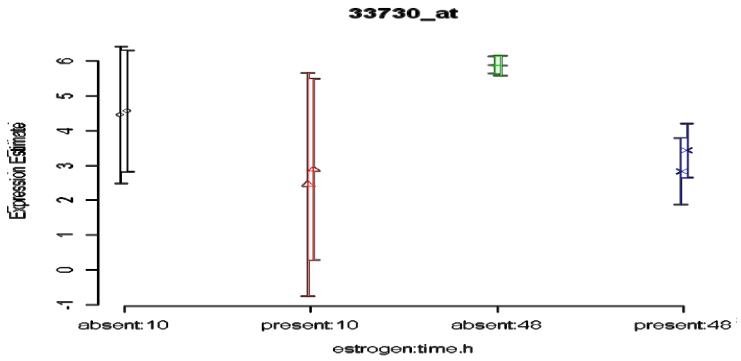


Probability of Positive Log ratio: PPLR

$$P(\mu_1 > \mu_2 | D, \phi) = \int_0^{\infty} P(\mu_1 - \mu_2 | D, \phi) d(\mu_1 - \mu_2)$$

Example:
In differential Expression analysis the goal is to estimate genes that change across conditions.

What happens then if we do not evaluate uncertainty?



Data from Choe et al, *Genome Biology*(2005)

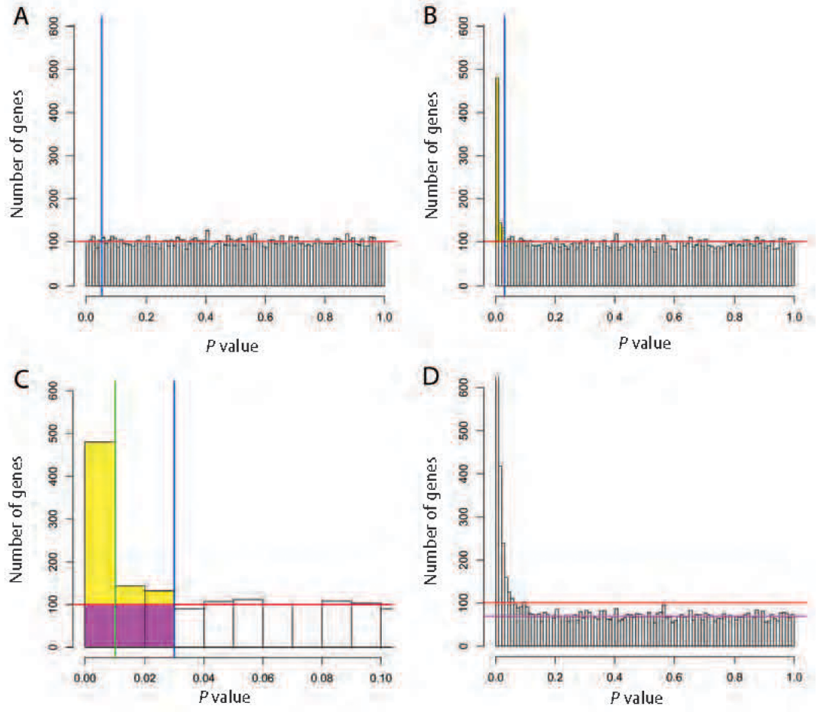
The University Of Sheffield.

The problem of Multiple sampling

small number of replicates due to the high cost of data generation, and the need for a very large number of significance tests.

10,000 genes

A. No DE
B. 500 DE
C. Zoom in
D. 1000 DE



how can we determine an appropriate significance level?
Methods to correct the significance level are called multiple tests corrections (or multiple comparisons corrections)

The University Of Sheffield.

Multiple testing

The **Bonferroni correction** is a classical correction method
the significance level to be used for each of 10 000 tests is 0.05/10 000 = 5 x 10-6.
The Bonferroni correction makes the probability of selecting one or more false positives equal to 0.05.

A fundamental principle of statistical analysis is that fewer false positives are associated with more false negatives (type II errors), that is, a higher chance of missing truly differentially expressed genes.

We need a more relaxed method than the Bonferroni correction.

However, one common problem associated with the classical type of multiple tests correction is that when the number of positives is relatively close to the expected number for false positives, we cannot estimate how many of the positives could be true positives.

The University Of Sheffield.

Then What do we do?

LIMMA: propose a moderated t-statistic
Smyth, G. K. (2004). Linear models and empirical Bayes
methods for assessing differential expression in microarray experiments.
Statistical Applications in Genetics and Molecular Biology 3, No. 1,

PPLR –*puma:* This model takes into account both the technical and the biological variance and significance of differential expression can then be evaluated by calculating probabilities.

$$FDR(1..n) = \frac{1}{n}\sum_{i=1}^{n} 1 - PPLR(i)$$

The University Of Sheffield.

False Discovery Rate

no correction for multiple hypotheses is too optimistic

Bonferroni's correction: too pessimistic…

An alternative is the false discovery rate (FDR).
FDR = number of false positive features/ number of significant features

false discovery rate (FDR) has become a standard for multiple tests correction in microarray data analysis. The idea of FDR put forward by Benjamini and Hochberg (Benjamini & Hochberg, 1995) is that it would be convenient to know what per cent of the positives discovered by multiple significance tests are false positives

NOTE: FDR depends on the distribution od p-value and not on the number of tests.

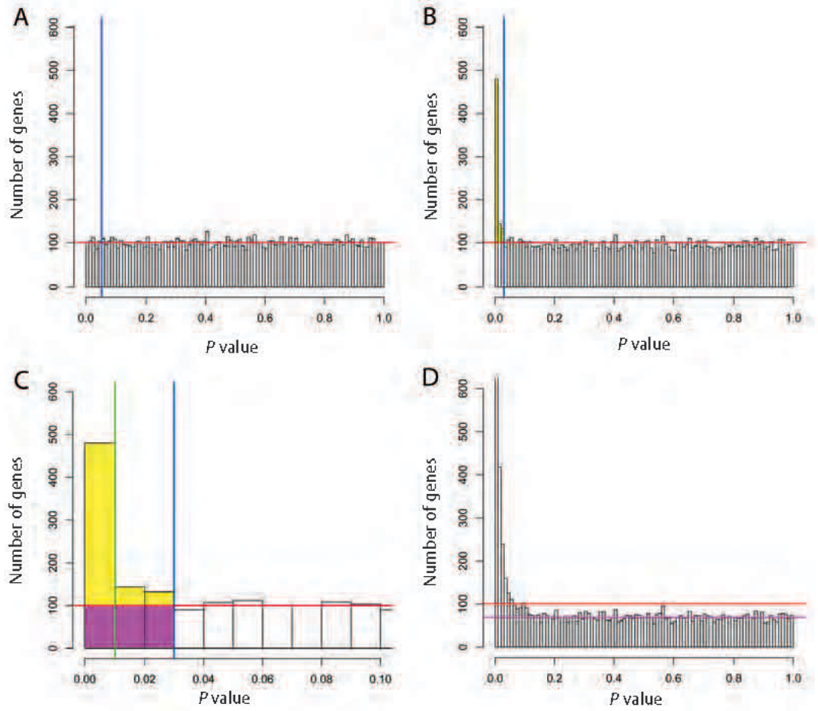
Storey also proposed the *q* value, which is defined for each gene as the minimum FDR that makes the gene a positive (Storey & Tibshirani, 2003).

False Discovery Rate

10,000 genes

A. No DE
B. 500 DE
C. Zoom in
D. 1000 DE

TP:456
FP:300



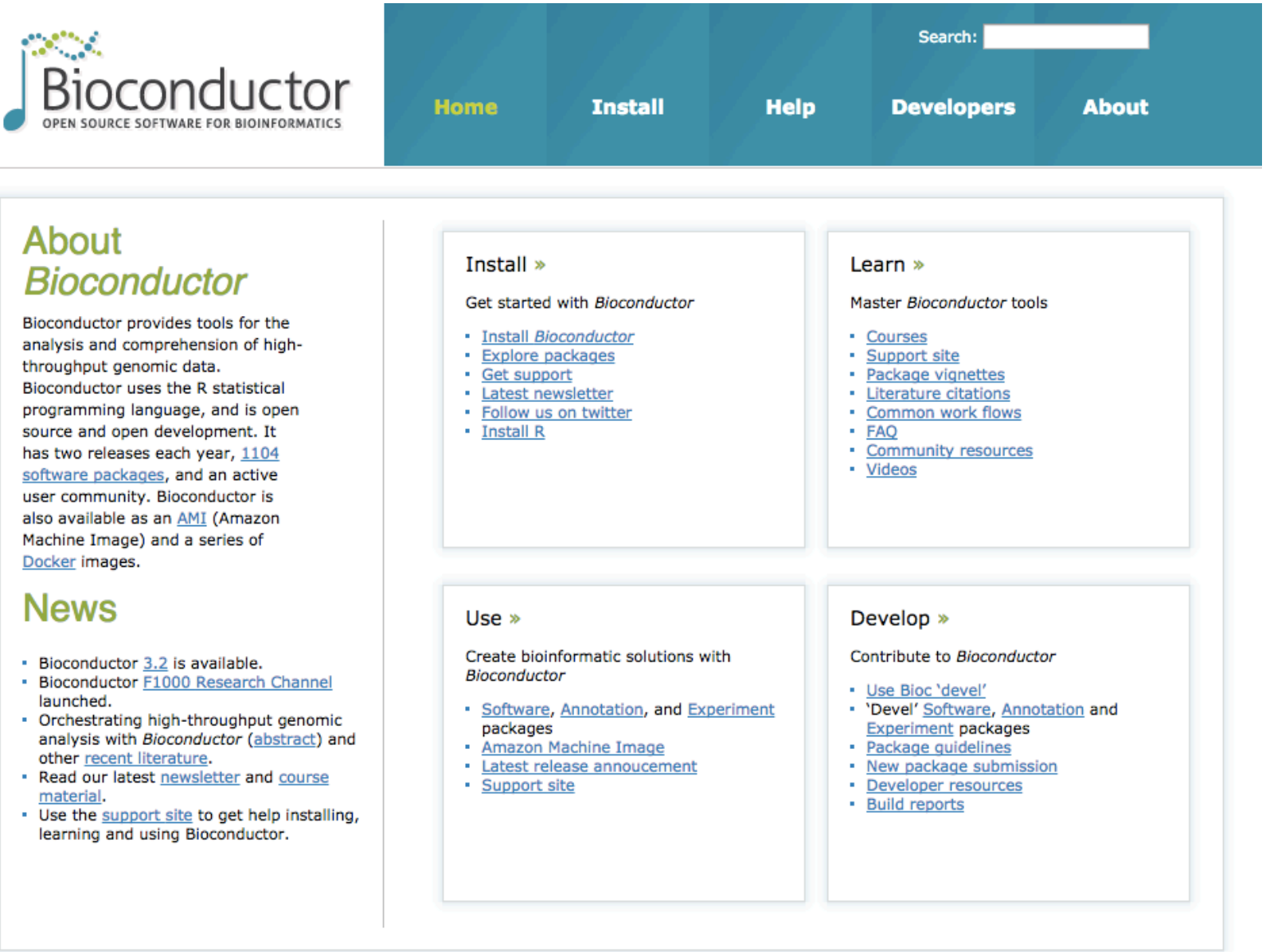
when a significance level of 0.03 is used, the FDR = 300/756 = 0.40. Now we know that on average 40% of 756 genes are false positives

if we use a significance level of 0.01 (green line), the number of true positives is 380, and the number of false positives is 100, so the FDR = 0.21

At a significance level of 0.0007321 the FDR = 0.05, which corresponds to (on average) 141.5 true positives and 7.5 false positives – with Bonferroni an average of 3 true positives and no false positives.

The University Of Sheffield.

Bioconductor



The University Of Sheffield.

Bioconductor (cont.)

- Bioconductor is an open source and open development software project, many packages are added and updated very frequently.
- Today we are just having a brief introduction into the usage and utilities of a small subset of packages from the Bioconductor project.
- Some of them are 'personal selection' based on what is my experience in teaching and research.
- To obtain a broad overview of available Bioconductor packages, it is strongly recommended to consult its official project site and remember always to check on updates.
- It is absolutely critical to use the original documentation of each package (PDF manual or vignette) as primary source of documentation.

The University Of Sheffield.

Bioconductor (cont.)

Affy Packages

These are the key packages for the analysis of affymetrix microarray data.

Affy

The Affy package provides the basic single point statistics methods for analysing affymetrix oligonucleotide arrays

Obtaining log transformed expression values with 3 different methods (MAS5, RMA, GCRMA)

Puma Packages

These are the packages for the analysis of affymetrix microarray data with probabilistic methods.

The gene expression levels are stored in what is called *Expression set* and the experimental design in what is called *Pheno Data*

The University Of Sheffield.

Bioconductor and limma

Visualization and quality controls

We will use hist(), boxplot() and mva.plot() and the *limma* version to visualise the data from the expression sets.

Analysis of Differentially Expressed Genes

Limma

Limma is a software package for the analysis of gene expression microarray data, that make use of linear models for analysing *designed experiments* and the assessment of differential expression.

The differential expression methods apply to all array platforms and treat Affymetrix, single channel and two channel experiments in a unified way.

Refer to the *limma* manual for examples and methods.

The University Of Sheffield.

Limma requires two matrices are specified:

- the design matrix which provides a representation of the different RNA targets which have been hybridized to the arrays.

- the contrast matrix which defined the combination of comparisons (*contrasts*). For very simple experiments the contrast matrix may not need to be specified explicitly.

FileName	Target
File1	WT
File2	WT
File3	Mu
File4	Mu
File5	Mu

```
> design

       WT MUvsWT
Array1  1      0
Array2  1      0
Array3  1      1
Array4  1      1
Array5  1      1
```

The University Of Sheffield.

How do we use limma?

Fist step :
define the matrices

Second step:
Fit a linear model to explain the relation between each gene on the array and the design matrix

$$E(y_i) = X\alpha_i$$

Where y_i is the gene expression for gene i, X is the design matrix and α_i is the coefficient for the gene i

Third Step: define significance levels for the comparisons. We use an empirical Bayes t-test, to calculate the p-values. This is to ensure that the information is learned from the data as a whole rather than single point.

Fourth Step: Select the significant targets and visualise the data

The University Of Sheffield.

Summary

We have seen today how to estimate gene expression

- single point statistics
- probabilistic models
- Uncertainty and how we can use it

Differential Expression Analysis

- p-values and their problems with gene expression data
- False Discovery Rate

Bioconductor Project

limma for Differential Expression